

MAKING THE INVISIBLE VISIBLE: CONTENT DEVELOPMENT BY DIGITIZATION OF EVAPORATING KNOWLEDGE FROM GUJARAT

P C SHAH

SUYASH KHEMKA

Abstract

This papers deal with the various aspects related to digitization. The digital library is the widely accepted term describing the use of digital technologies to acquire, store, preserve and provide access to information and material originally published in digital form or digitized from the existing print, audio-visual and other forms. The authors are discussing their plan to digitize rare and out of print books published before 1900 and available in over 100+ libraries in Gujarat and to digitize grey literature of research value published by about 150 big and small R&D institutions and NGOs in Gujarat in all subject areas.

Keywords: Digitization/ Rare books/ Grey literature/ Evaporating Knowledge/ Gujarat/ Lightening

1. Introduction

There are many purposes and ways to create the digitized images, depending upon the uses, needs and users demand. Digitization assures high quality of information of wide varieties of resources to wide varieties of users, possible any time, any where simultaneously. However, digitization is used and advocated for:-

1. Preservation of quality.
2. Multiple referencing. Access of information to several users at a time.
3. Accessible to users at any distance and at any time.
4. Security Measures.
5. Restoration of rare and out of print material i.e. preserve and conserve material.
6. Increasing their customer base by making older and rare or heavily used documents available to a much wider audience.
7. To provide quality services in a difficult but hopeful environment.
8. It enables change in the individual & group behavior. Hence, it is a powerful instrument of change.
9. It evolves into an active medium of communication which effect social & work practices.

In short digital documents have the ability to enable world-wide access to a never ending supply of distributed information which is constantly and conveniently available and updateable.

2. Libraries as Memory Institutions

Libraries of the world exist to hold and preserve the human documentary heritage in the human documentary heritage in whatever form that heritage might be represented and upon whatever medium it might be inscribed, and to make it available to a wide range of users.

Libraries work as a “memory institution” with a responsibility for maintaining human kind’s recorded memory. “Archives, libraries and museums are memory institutions. Their collection, contain the memory of peoples, communities, institutions and individuals, the scientific and cultural heritage, and the products throughout time of our imagination, craft and learning. They just join us to our ancestors and are our legacy to future generations. They are used by the child, the scholar, and the citizen, by the business person, the tourist and the learner. These in turn are creating the heritage of the future. These memory institutions contribute directly and indirectly to prosperity through support for learning, commerce, tourism and personal fulfillment” (Dempsey L. (1999), Scientific, industrial & cultural heritage; a shared approach, New Review of Information and Library Research, 5, 3-29).

From their inception Libraries have provided gateways to knowledge, have reflected the plurality and diversity of society and have supported the process of democratization. Information is power, and ensuring access to information is the role of the librarian, now more than ever.

3. Coverage and Scope

In this article we are discussing our plan

1. To digitize rare and out of print books published before 1900 and available in over 100+ libraries in Gujarat. (Manuscripts will not be covered in this project)
2. To digitize grey literature of research value published by about 150 big and small R&D institutions and NGOs in Gujarat in all subject areas.

4. Benefits of Digitization of Rare & out of Print Books and Grey Literature

- a. It will democratize the knowledge available in such thousands of resources.
 - b. All people over the world will get the needed information quickly. It does not matter if you have it; it only matters if you can get it.
 - c. Digitization of such documents will bring to libraries the possibility to extend their reach beyond the traditional user base.
-

- d. All digital information will be portable i.e. all the required literature can be downloaded at a time and anywhere.
- e. Efficient: It will eliminate cost to all for storage, distribution and transport as all the digitized database will be stored at one place called "Gujarat rare & out of print books and grey literature information centre".
- f. Easily searchable.
- g. Far distance people will know about the availability of such literature. Hence, wide base of users.

5. The Economic Advantages of Digital Information

1. The Communication Effect: An increased amount of information is exchangeable at a reduced cost.
2. Access: The digital format allows multiple access to a single copy, thus offering economies of scale in access and delivery; remote access to resources offers users the benefit of not having to travel to the resource.
3. Flexibility: The same digital resource can be used for multiple purposes and for many different classes of user without the need to multiply the cost of its creation.
4. E-Commerce: Digital resources enables amore effective match between the buyer and seller.
5. Product Costs: The creation chain for new publication is generally digital first and then output to analogue format; digital information is "Cheaper to produce, store, distribute and reproduce or copy".
6. Volume Enhancement: Generally the greater the volume of digital information the greater the value placed upon the resource or service.
7. Availability of services: Information accessible via a digital library is usually available at all, without related to library opening times or available staffing.

6. Digitization of rare and out of print books

It is known that Gujarat has a rich heritage and pioneer in starting systematic library movement in India. Gujarat has about 2000 libraries and Information Centres (LIC) as per 'Directory of Libraries and Information Centres in Gujarat' published by Gujarat Granthalaya Seva Sangh (GGSS)', Ahmedabad. Out of these 2000, there are more than 70 LIC which have celebrated their century and there are many in pipeline to celebrate century. Over and above there are many LIC who have rare and out of print books published before 1900. All these will be covered in this project.

Year 1900 is selected as Copyright Act is not applicable.

These rare & out of print books are deteriorating day by day and at a fast rate. If we do not awake urgently then they will evaporate (destroyed) either by earthquake, water, theft or natural deterioration or for any other reason, as it happened in many LIC in Bhuj-Kutch in 2001 during earth-quake. One of the authors (P.C. Shah) of this paper, in 2001 surveyed the earthquake affected 45 libraries. He found that in about 20 public libraries and few museums have lost their valuable works & other rare collection which were published before 1900.

Over and above far distance people have no idea about availability of such rare and out of print collection in particular LIC.

Unfortunately these LIC have no funds, manpower and technology to preserve and utilize these evaporating knowledge contained in books in their own library.

We have an economically viable plan to digitalize such collection and to put their bibliographical information / details on WEB so that it is known all over the world. Interested users can contact us for getting full text or for further information on its availability, etc.

Estimated number of unique titles

Our primary survey indicates that there are over 15,000 unique titles which are published before 1900 and are available in around 100 LIC in Gujarat. Duplication in all ways will be avoided.

7. Digitization of Grey Literature

Another type of collection in Gujarat which needs digitization as early as possible is the 'Grey Literature' published by over 150 R&D institutions and many NGOs in Gujarat in various subject areas. Many of them are over 50 years old. Since their inception, they publish research reports, Circular reports, Committees reports, Vision documents, Feasibility reports etc. These are not available through normal channels of distribution such as book sellers, publishers, etc. These are produced for specific purpose and submitted either to sponsors or to member industries / companies or members of the committee. Contents and data contained in these reports are rare and have high research value. Limited number of copies are produced. Many a times copies of such publications are not available even in the institutions for historical record and for further extension of research in the same subject in future. Realizing that such publications have further research value to others (as Old is Gold), we plan to digitize the grey literature available in R&D institutions & NGOs in Gujarat.

Further the problem of development of R&D is being global concern, there is a need to capture information and full text of grey literature and rare and out of print books and disseminate the same to International communities who are interested in such research / content of rare R&D literature. These types of literatures may become a basic work for further research in particular areas or sub areas (subjects).

Estimated number of documents in grey literature

It is estimated that there are about 30000 grey literature reports on various subjects in various R&D institutions & NGOs in Gujarat. All these are not available / known to the world.

Further, it is very difficult, if, not impossible to obtain 'Grey Literature', as such literature seldom come under bibliographical control.

8. Objectives of the Project

In view of above characteristics, objectives of this project are:

- a. Lightening the literature making the invisible visible.
- b. To find ways and means of enhancing the accessibility and availability of rare and out of print books and unpublished grey literature (documents) through bibliographical and full text through document delivery service in digital form or in hard copy form as per users requirements.
- c. To digitize full text of rare books & grey literature.
- d. To build up strong document collection in digital form, of rare and out of print books available in Gujarat and grey literature published in Gujarat. Thereby GGSS will act as a "Gujarat Rare Books & Grey Literature Information Centre" and hence ensure the availability of such evaporating knowledge of Gujarat to the world.
- e. To alert and bring to the notice of world over users, the information about rare books and R&D works of research institutes & NGOs, etc. of Gujarat.
- f. To increase the reference value of rare and out of print books & grey literature available in Gujarat.
- g. To increase the utility of rare books & researches done in Gujarat. In turn research work of R&D institutions will be further known world over and thereby these R&D institutions will earn more reputation and may get sponsored projects in their respective areas. This is an indirect benefit to R&D institutions & country.
- h. To provide in digital form / hard copy form the full text of rare books and out of print books & grey literature at a reasonable cost to users after taking due permission from the originator.

This will earn some money to both: to proposed information centre and producer / originator of such literature.

9. Organization and Action Plan

Careful planning of digitization work would bring clarity to this project. It would require inputs from diverse branches of knowledge.

9.1 Feasibility of the work

Such rare and out of print books and grey literature are scattered over many libraries and institutions. Further, no body will trust on individual or unknown organization. It may not be economically viable for each library or R&D institutions to digitize these rare & out of print books and grey literature using world latest technology.

Such work requires trustworthiness, better knowledge of local libraries and institutions / NGOs, expertise in latest technology and finance. Hence, such work can be done comprehensively and in better way and economically by trustworthy local bodies. Hence, only known and trusted local voluntary academic trust (agency) like Gujarat Granthalaya Seva Sangh (see website: www.ggss.org.in) can take and fulfill such works quickly and faithfully with the help of experts.

9.2 Human Resources Planning

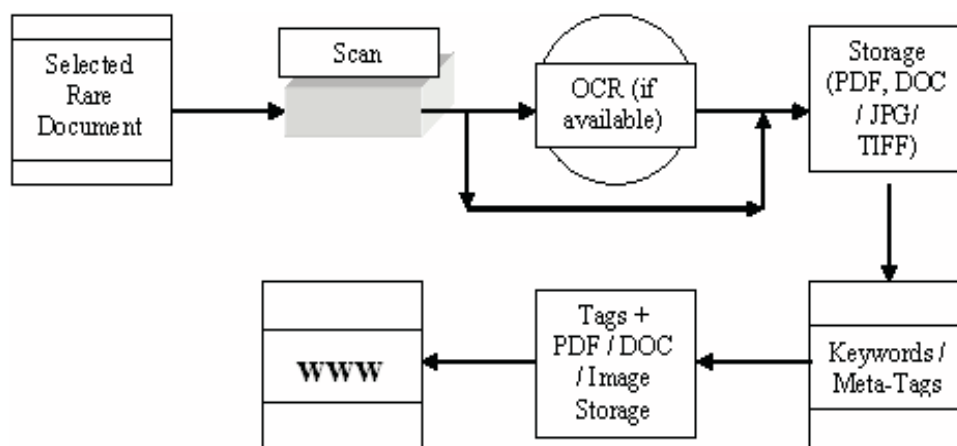
Two authors of this paper will take full responsibility of implementation and completion of this work. Suitable manpower will be trained as per requirements and will be paid accordingly. Where ever possible outsourcing will be done in order to economize on expenditure and to take best advantage of experts in the areas.

9.3 Managerial Planning and implementation

- One of the author, Shri P. C. Shah is President of Gujarat Granthalaya Seva Sangh, Ahmedabad, which is a Government recognized State Library Association. He has seen majority of the libraries and R&D institutions in Gujarat and know majority of the trustees, libraries & librarians personally. He has been associated / known amongst many NGOs. He has more than 45 years of experience in the field of LIS. Hence, special letters will be written to all identified over 100 LIC having rare and out of print books, and general circular regarding this project to all members of GGSS alongwith its quarterly publication 'GRANTHLOK'. Second author is an experienced IT expert and MBA (Marketing).
 - Over and above 'Letter to the Editor' will be sent to various 'Daily Newspapers' published in Gujarat informing all about this project. This will help in getting information from all other sources regarding rare & out of print books available in Gujarat and grey literature published in Gujarat.
 - Director of Libraries will be requested to issue a circular to support this project and to provide required facilities & information, etc.
 - After receiving, information on number of rare and out print books available in these LIC and grey literature in various R&D institutions and NGOs and after analyzing the received data, one day workshop of Trustees & Librarians of LIC and R&D institutions including NGOs will be organized to inform them the details of the project and about the benefit they will get by cooperating in this work. Detail plan for execution will be prepared in consultation of these LIC and R&D institutions & NGOs.
-

- Complete digitization work will be executed at place of LIC and at various R&D institutions, etc.
- A copy each of the digitized work (in the form of CD) will be given to that particular LIC & Institutions at free of cost.
- We will also help / advise them in preserving the originals.
- Annotated bibliographical details will be hosted on our GGSS website so that people in India and abroad can know about such rare and out of print books and grey literature available in Gujarat. Hence, there will be maximum exposure and utilization of rare and out of print resources (knowledge) and grey literature.
- Rare & out of print books will be in English, Hindi, Gujarati & Sanskrit and grey literature may be in English and also in Hindi & Gujarati. Hence bibliographical details will be given in Roman Scripts.

10. Digitization Process



Schematic Diagram of Digitization Process

11. Technologies

The process of digitization seems very simple though a wide set of sophisticated tools are available to execute the task. The scanned document may be either stored as image, image is nothing but a composition of grid of pixels, or text file using OCR which can enable full-text search and transliteration of the scanned text in any other language. Below we are trying to represent an in-depth overview of each of these technologies that is of great benefit in this digitization process.

Scanning

Capturing a digital image is known as scanning. The scanning process can provide the output either in image form or can convert the same into text using OCR. Generally

the image scanned is measured in terms of pixels (size) and resolution (dpi, dot per inch). The better the resolution the better the quality of the image and the greater the file size of the image.

There are different types of devices available in the market to capture the digital image primarily divided into two Scanners & Digital Cameras. Scanners are the most preferred and cost effective solution for digitization of documents.

Following type of scanners may be used for digitalization process:-

1. Flat bed scanners
2. Book scanners
3. Robot scanners

The extended details of each of these are discussed in the “Equipment Required” section of this document.

File Formats

Another related topic to scanning is file formats. There are various formats available to store the scanned document. There is relatively very less difference amongst each of them. The images scanned can be stored in either GIF, JPEG or Tiff format while PDF is the most preferred format for textual scanned documents. The difference type of storage primarily helps to handle the compression ratio of the scanned image without affecting much on the quality aspect. Following table would give you an idea about the different file formats available to us:

Format	Encoding	Compression	Quality	Portability	Origin
GIF CompuServe Graphic InterChange Format CompuServe	Binary	LZW	8 bits	Mac/ PC/ Unix	
JPEG / JPG Joint Microsystems Photographic Expert Group	Binary	RLE & JPEG	24 bits	Mac/ PC/ Unix	C-Cube
TIFF	Binary	CCITT Gr. 3 & 4 LZW, RLE, JPEG	24 bits	Mac / PC/ Unix	ALDUS & Microsoft
PDF	ASCII & Binary	None Recently added JPEG	32 bits	Platform Independent	Adobe System

OCR – Optical Character Recognition

Another interesting and relatively useful technology in digitization is OCR – Optical Character Recognition. When we scan any document it is converted into image. Using OCR this image can be converted back into the text format which can be edited if desired. Further this converted text can also be of great help in digitization by providing full-text search to the user. Currently, OCR is available for English at global level while recently CDAC and Government of India have released OCRs for Hindi & Tamil as well. As such this restricts the use of OCR only to these three languages.

Meta Tags / Data

Any digitized data that is stored needs to be retrieved. In order to retrieve and locate the same it is very much essential to define Meta Tags or Meta Data. The information content in this provides the data with respect to the encoded / attached file. The “Dublin Core” (see http://purl.oclc.org/metadata/dublin_core/) and other special initiatives for structuring and standardizing the description of the data propose to include information about the characteristics of the digital files attached. The combined information including location and summary of their contents is known as Meta Tags / Meta Data.

PDF

Another alternative to tagging is the use of proprietary format such as Adobe Portable Document Format (PDF) which is the open standard for electronic document distribution worldwide. Consisting of a package of software, PDF is capable to handle scanning, OCR conversion and structuring of text and images.

12. Equipment / Softwares Required: Alternative & Selection Why? :

Computers / Storage Server

The equipments required for processing / storing the digitized document would be Computer Systems and Storage Server. Storage servers are basically systems having huge storage capacity and multiple storage devices connected to it including CDRW, DAT Drive, Multiple Terra Bytes of Disk Space etc.....

Scanner

In digitization process we have got option of three options available to us for scanning the document:

Flatbed Scanner : The most common scanner available in the market. This scanner is highly cost effective. While cost and availability can be put as its benefit, its slow speed and inability / poor scanning at the spines and edges of the book have to be considered as the disadvantage.

Images are placed on a glass bed either with or without a holder. The scan area varies in size from 8-1/2" x 11" to 13" x 18". Either the bed is stationary and the scanning head moves or if the bed moves, the scanning head is stationary. They are either a single-pass or three-pass scanner. Single-pass captures all the RGB colors by moving the light source over the image once. Three pass scanners use three passes, one pass each for red, green and blue. The single-pass scanners are faster but the three-pass scanners are generally more accurate.

Flatbed scanners can scan originals of varying thicknesses, and some are capable of scanning three-dimensional objects. You can add adapters for automatic page feeders. There are also templates you can use to hold pieces such as transparencies or slides.

In traditional flatbed scanners, the scanning head moves in one direction only. There is a new technology called XY scanning which positions its scanning head along an XY axis. The scanner head slides both horizontally and vertically beneath the bed. The XY scanning technology assures high resolution and uniform sharpness of the entire scanning area. It also makes it possible to enlarge an image to a much higher percentage than the traditional flatbed.

The highest resolution you can achieve without interpolation is about 5,000 dpi. With interpolation, the resolution may increase to about 11,000 dpi.

Book Scanner : Second best option available in the market is book scanner. This scanner is comparatively lower in rate and speed is also acceptable. But the main problem is manual handling of the document takes huge time. This scanner becomes mandatory in cases where we have documents which have become brittle in nature and needs careful handling.

The Book Scanner has revolutionized the way that information found in books, newspapers, delicate over-sized documents, and 3-D objects is captured. Book Scanner makes scanning in color, black and white, or grayscale a snap. You can even choose to scan once and create both color and grayscale or other output combinations.

Using Book Scanner



Using Flat bed Scanner



Book Scanner overhead scan design and scan/print electronics with auto-focus and book-fold correction not only prevent damage to delicate books, but also make using Book Scanner easier and more efficient than normal scanners. With powerful functions like automatic removal of unsightly borders and Automatic Document Location (ADL), Book Scanner has redefined the industry standards.

Robot Scanner : The third and the best option for digitization is Robot Scanner. Robot scanner has got excellent speed and a robot arm which flips the pages on its own. All this while keeping excellent quality. The only drawback in this case is its high cost and inability to handle soft and brittle documents.

	Robot Scanner	Book Scanner
Throughput Preservation prepping,	Up to 2,400 pages per hour Scan bound books without straining even fragile bindings	100-150 pages per hour Requires extensive and aggressive page flattening
Quality	300-600 dpi, full-color images without distortion from the first page to the last	Variable
Ease of Use	Simple, GUI (Graphical User Interface) minimizes training	Variable

Softwares

The main component in the whole digitization process is the software used to store and retrieve the digitized document. With advent of open source era we have got three options in this space:-

Greenstone : Greenstone is a suite of software for building and distributing digital library collections. It is not a digital library but a tool for building digital libraries. It provides a new way of organizing information and publishing it on the Internet in the form of a fully-searchable, metadata-driven digital library. It has been developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium. It is open-source, multilingual software, issued under the terms of the GNU General Public License. Its developers received the 2004 IFIP Namur award for “contributions to the awareness of social implications of information technology, and the need for an holistic approach in the use of information technology that takes account of social implications.”

GDL : GDL stands for GDL Digital Library. Originally, G here means “Ganesha”. It is a name of digital library application that is developed by KMRG ITB.

Lead by Ismail Fahmi (ismail@itb.ac.id) at Knowledge Management Research Group (KMRG) kmr@kmr.itb.ac.id, a free software for Digital Library has been developed. A

working network of the Indonesia Digital Library can be found at <http://www.indonesiadln.org> and <http://gdlhub.indonesiadln.org>. The International Development Research Center (IDRC, <http://www.idrc.ca>) Canada and Yayasan Litbang Telekomunikasi Informatika (YLTl, <http://www.ylti.or.id>) funded their initial research activities.

The Ganesha Digital Library, an open-sources/free software, can be run on a system running Unix/Linux or Windows 98/NT/2000 or Windows XP with winsock32. Apache web server using PHP4 scripting language to interact with MySQL database may be used by dedicated or dial-up infrastructure to create the knowledge infrastructure.

DSpace : DSpace is a groundbreaking digital repository system that captures, stores, indexes, preserves, and redistributes an organization's research data.

Jointly developed by MIT Libraries and Hewlett-Packard Labs, the DSpace software platform serves a variety of digital archiving needs.

13. Cost & Sources of Finance

The total cost of the project is estimated at Rs. 40-50 lacs depending upon the equipments selected by the experts.

We are hopeful to get finance for the implementation of this project from various agencies like Ministry of HRD, Government of India, NGOs, Gujarat Government, RRLF, Contribution from various donors, etc. Over and above:

1. Full text of digitalized documents will be available at cost from Information Centre established by GGSS to users anywhere in the world. The income will be suitably shared between concerned LIC / institution and Information Centre of GGSS.
2. Income will help GGSS to update the site and to modify if needed. Income will also help for digitization of more & more out of print documents and grey literature in future.
3. In future proposed 'Digital Information Centre' will also take digitization work from outside agencies (Libraries, Institutions, etc.). Hence, there will be maximum utilization of equipment & other resources as well as expertise.

14. Conclusion

Digitization of Knowledge contained in rare and out of print books and researched knowledge hidden in grey literature spread all over Gujarat will be the unique work. This will make the invisible knowledge visible to the world and expand the user base of such knowledge. It would be an excellent and unusual content development in Digital Forms of evaporations knowledge from Gujarat. Sooner we awake and save before everything – such knowledge is destroyed – (died) – because of earthquake or for any other reason. Sooner the better we awake and save our golden heritage.
